Melanie Green, Miriam Ayafor, and Gabriel Ozon (2016), *A spoken corpus of Cameroon Pidgin English: pilot study.* University of Oxford Text Archive.

Research Purpose and Scope

Cameroon Pidgin English (CPE), a primarily unwritten language, is a typical pidgin/creole language in that it has a European colonial language as one of its ancestors, but also has characteristics of West African languages. This project to establish a 240,000-word digital corpus was enabled by a British Academy/Leverhulme small grant (ref. SG140663) of £9,857 for which Green was the PI and made a 40% contribution to the project. The corpus consists of recordings from over 80 speakers, together with full transcripts and a word list. A pioneering pilot study, it provides the foundation for a planned one-million-word corpus of CPE, the size at which corpora yield reliable information about lexical phenomena. As a substantial study in its own right, the corpus significantly augments digital humanities practice and advances linguistic research in three principal ways:

- It provides the first spoken corpus of a pidgin/creole language, offering a unique resource for linguistic research, as well research on Cameroonian culture and society more generally.
- It offers a methodological model for the development of corpora of other unwritten languages, resources essential to our understanding of the world's languages.
- It supported the training of a Cameroonian linguist in corpus-building, skills that are desperately needed in one of the most linguistically diverse regions of the world.

Description

This resource is a 240,000-word corpus of spoken Cameroon Pidgin English (CPE), a widely-used yet stigmatised and largely uncodified pidgin/creole variety.

The corpus consists of transcriptions of private and public dialogues and monologues, with mark-up and POS-tagging, together with accompanying sound files. The recordings were conducted in five different locations in Cameroon (Bamenda, Buea, Douala, Kumba and Yaounde), allowing some insights into regional variation. Text categories and the proportions of monologue and dialogue are guided by those of the International Corpus of English (ICE) project, which makes the corpus immediately comparable with existing corpora of post-colonial varieties of English.

- **Spelling**: since there is no standardised orthography for CPE, the orthography adopted for this project is based on that developed by Ayafor (2014), which was kept under review during the course of the project.
- **Annotation** was added to the transcriptions based on ICE guidelines for the annotation of spoken texts: standard mark-up symbols were used to denote text unit, speaker

identification, overlapping speech, unclear words, uncertain transcriptions, anthropo-phonics, editorial comments, foreign words and indigenous language words.

- **Tagging**: a parts of speech (POS) tagset for CPE was devised based on CLAWS 5. Initially tagging was conducted manually, and then by means of TreeTagger. A third of the corpus has been post-checked, with accuracy rates at 94%.

The corpus is aimed at providing a resource for linguistic description and comparison. It allows linguists to identify and describe recurring grammatical patterns, as well as the phonology of the language (given the availability of sound files deposited with the text files). It also allows comparison of CPE with other pidgin/creole languages, other Cameroonian and West African languages, and other varieties of post-colonial English. Furthermore, the corpus provides an exceptional resource for the study of general/theoretical linguistics, creolistics, typology, language contact and change, sociolinguistics and discourse analysis.

The corpus contains 80 sound recordings of monologues (scripted and unscripted) and dialogues (public and private). Each sound file (in .wav format) is 10-15 minutes in length. These recordings have been transcribed (each approximately 3,000 words in length) and annotated. Transcriptions are submitted in two formats: (a) plain transcription (with basic markup indicating speaker turns, overlaps, etc.), and (b) a POS-tagged version, which adds POS-tags to the plain version of the transcription.

The language of the monologues is Cameroon Pidgin English, with codeswitching into English, French, and indigenous Cameroonian languages.

Dissemination/Access

The significance and scale of this project is demonstrated in the further scholarship it has generated. It is introduced and discussed in Gabriel Ozón, Miriam Ayafor, Melanie Green and Sarah FitzGerald, 'The spoken corpus of Cameroon Pidgin English', *World Englishes* 36.3 (2017), 427-47. Research for this project subsequently underpinned Miriam Ayafor and Melanie Green, *Cameroon Pidgin English: A Comprehensive Grammar* (Amsterdam: John Benjamins [London Oriental and African Language Library], 2017); Melanie Green and Gabriel Ozón, 'Information structure in a spoken corpus of Cameroon Pidgin English' in Evangelia Adamou, Katharina Haude and Martine Vanhove eds, *Information structure in lesser-described languages: studies in prosody and syntax* (Amsterdam: John Benjamins, 2017), pp. 329-56; and Melanie Green and Gabriel Ozón, 'Valency and Transitivity in a Contact Variety: The Evidence from Cameroon Pidgin English', *Journal of Language Contact* 12.1 (2017), 52-88.

The corpus (sound files, transcriptions/annotations, and metadata) was completed in 2016 and is hosted by the Oxford Text Archive at the Bodleian Library, University of Oxford. It is publicly available under an Attribution-NonCommerical-ShareAlike 3.0 Unported (CC BY-NC-SA 3.0) license:

https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/2563#